

Color-Based Object Tracking in Multi-Camera Environments

In Proceedings of the DAGM'03, Springer LNCS 2781, pp. 591-599, Sep 2003

Katja Nummiaro¹, Esther Koller-Meier², Tomáš Svoboda²,
Daniel Roth², and Luc Van Gool^{1,2}

¹ Katholieke Universiteit Leuven, ESAT/VISICS, Belgium
{knummiar, vangool}@esat.kuleuven.ac.be,

² Swiss Federal Institute of Technology (ETH), D-ITET/BIWI, Switzerland
{ebmeier, svoboda, vangool}@vision.ee.ethz.ch

Abstract. Smart rooms provide challenging research fields for surveillance, human-computer interfacing, video conferencing, industrial monitoring or service and training applications. This paper presents our efforts toward building such an intelligent environment with a calibrated multi camera setup. For a virtual classroom application, the best view is selected automatically from different cameras that follow the target. Real-time object tracking, which is needed to achieve this goal, is implemented by means of color-based particle filtering. The use of multiple models for a target results in a robust tracking even when the view on the target changes considerably like from the front to the back. Information is shared between the cameras and used for (re)initializing an object with epipolar geometry and the prior knowledge of the target model characteristics. Experiments in our research environment show the possible uses of the proposed system.

1 Introduction

Intelligent environments provide a wide variety of research topics like object tracking, face/gesture recognition or speech analysis. Such multimodal sensor systems attract application areas like surveillance, human-computer interfacing, video conferencing, industrial monitoring or service and training tasks.

Within this paper we focus on the autonomous processing of visual information based on a network of calibrated cameras. Each camera system comprises a recognition and tracking module which locates the target in the observed scene. Both modules operate on color distributions while multiple color models for a target are handled simultaneously. To document interesting events on-line, an automated virtual editor is included in a central server and produces a video stream by switching the camera to the best view in the room. Figure 1 illustrates the system architecture of our multi-camera setup.

For the integration of multiple cameras, the ViRoom (Visual Room) system by Doubek *et al.* [5] is used. The modular architecture is constructed from low-cost digital cameras and standard computers running under Linux and allows

consistent, synchronized image acquisition. We extended the ViRoom software by an automatic camera control to keep the target in the center of the view as various applications are more interested in a clear image view than in the exact location of the target. Accordingly, from the detected location, the pan and tilt angles are calculated and the cameras are rotated into the appropriate direction. A color adjustment is applied during the calibration of the cameras since the images show in general different brightness and color characteristics.

The research area of multi sensor systems is very active [3, 8, 9, 13]. For instance, a flexible multi-camera system for low bandwidth communication is presented by Comaniciu *et al.* [3]. Based on color tracking the target on the current image can be transmitted in real-time with high resolution. Khan *et al.* [8] describe an interesting approach to track people with multiple cameras that are uncalibrated. When a person enters the field of view of one camera, the system searches for a corresponding target in all other cameras by using previously compiled field of view lines. Krumm *et al.* [9] describe the tracking in an intelligent environment using two calibrated stereo cameras that provide both depth and color information. Each measurement from a camera is transformed into a common world coordinate system and submitted to a central tracking module.

The work most closely related to ours is that of Trivedi *et al.* [13], where an overall system specification for an intelligent room is given. The 3D tracking module operates with multiple cameras and maintains a Kalman filter for each object in the scene. In comparison we use a more general representation of the probability distribution of the object state which allows to initialize this distribution along the epipolar lines when an object enters the field of view of a camera. In our system the best view is selected according to the quality of the tracking results for the individual cameras while Trivedi *et al.* utilize the motion of the tracked target.

The outline of this paper is as follows. Section 2 presents a short review of the color-based tracking technique. In Section 3 the multiple target models used for the tracking are explained. Section 4 presents the exchange of information in the camera network while Section 5 explains the selection of the optimal camera view. In Section 6 some experimental results are presented and finally, Section 7 concludes the paper.

2 Tracking

Robust real-time tracking of non-rigid objects is a challenging task. Color histograms provide an efficient feature for this kind of tracking problems as they are robust to partial occlusion, are rotation and scale invariant and computationally efficient. The fusion of such color distributions with particle filters provides an efficient and robust tracker in case of clutter and occlusion. Particle filters [6] can namely represent non-linear problems and non-Gaussian densities by propagating multiple alternate hypotheses simultaneously.

The color-based particle filter [10, 11] approximates the posterior density by a set of weighted random samples $\{(\mathbf{s}_t^{(n)}, \pi_t^{(n)})\}_{n=1}^N$ conditioned on the past ob-

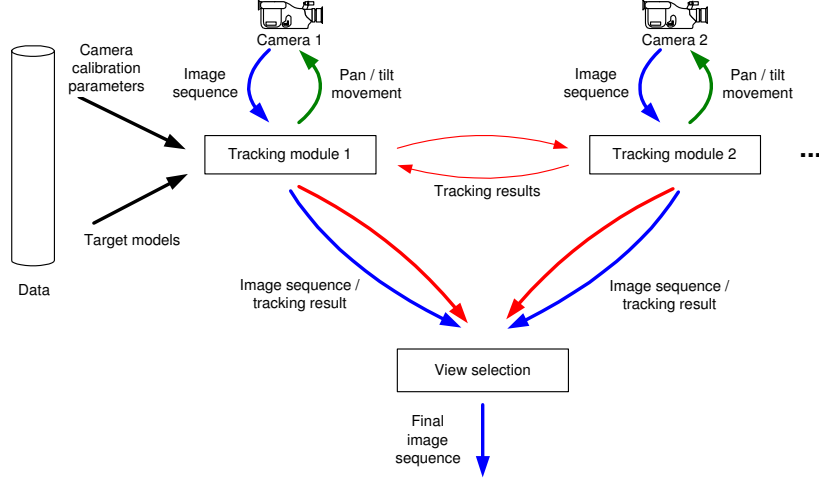


Fig. 1. The sketch of the system architecture with multiple cameras and the virtual editor.

servations. Each sample \mathbf{s} represents one hypothetical state of the object, with a corresponding discrete sampling probability π , where $\sum_{n=1}^N \pi^{(n)} = 1$. The tracked object state is specified by an elliptical region

$$\mathbf{s} = \{x, y, \dot{x}, \dot{y}, H_x, H_y, \dot{H}\} \quad (1)$$

where x, y represent the location of the ellipse, \dot{x}, \dot{y} the motion, H_x, H_y the length of the half axes and \dot{H} the corresponding scale change.

In order to compare the histogram of such a hypothesized region $p_{\mathbf{s}^{(n)}}$ with the target histogram q from an initial model, a similarity measure based on the Bhattacharyya coefficient [4, 1, 7]

$$\rho[p_{\mathbf{s}_t^{(n)}}, q] = \sum_{u=1}^m \sqrt{p_{\mathbf{s}_t^{(n)}}^{(u)} q^{(u)}} \quad (2)$$

is used, where u represents the number of bins for the histograms.

3 Multiple Target Models

To support multiple cameras, we have to use more than one histogram for a target, as for instance one camera might have a frontal view of the head, while an oppositely placed camera will have a view of the back of the head. Since we track independently in each camera, the correspondence between the histograms must be established.

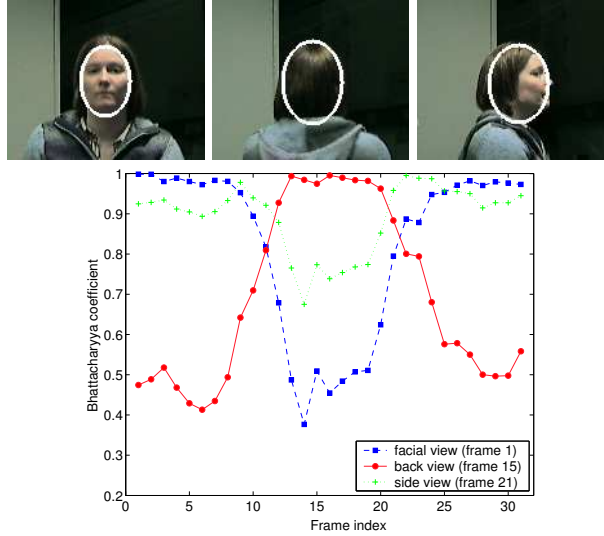


Fig. 2. Top row: The three main views and histogram regions for the target model (frames 1, 15 and 21). Bottom row: The plotted Bhattacharyya values from a turning head sequence, using the different target models (facial, side and back view) and comparing them to the head region.

Three characteristic images from the front, the side and the back are selected as initial target models from a recorded sequence of the person to be tracked and the corresponding histograms $q = \{q_f, q_s, q_b\}$ are stored. During the tracking, the similarity measures to these three histograms are included in the object state. By using a linear stochastic model for the propagation, the Bhattacharyya coefficients for the next frames can be estimated and rapid changes of this coefficient are restricted.

Figure 2 shows the different values during the turning head sequence that is used to determine the target histograms. As can be seen, the Bhattacharyya coefficients change smoothly.

The initial samples of the individual particle filters are spread over the whole image or strategically placed at positions where the target is expected to appear. A target is now recognized on the basis of the three Bhattacharyya coefficients where the best matching model is taken as the target model. By calculating the mean value μ and the standard deviation σ of the Bhattacharyya coefficient for elliptic regions over all the positions of the background in the initialization step, we define an appearance condition as

$$\rho[p_{s_t^{(n)}}, q] > \mu + 2\sigma. \quad (3)$$

This indicates a 95% confidence that a sample does not belong to the background. If a fraction $b \cdot N$ of the samples shows a high enough correspondence to one of the

target histograms the object is considered to be found and the tracking process is started. The parameter $b = 0.1$ has been proven sufficient in our experiments and is called the ‘kick-off fraction’. During tracking, the target model of each camera is adapted as described in [10].

4 Exchanging Information across Cameras

Exchanging information between the cameras is important to provide a robust tracking. Within our system we focus on the (re)initialization of the individual trackers. As we are working with a calibrated camera setup [12], we utilize the tracking results of specific cameras to initialize the remaining trackers via epipolar geometry. The epipolar geometry is used in two different ways: 1) During initialization when one of the target models matches an object, 2) When the object is temporally lost due to clutter, occlusions or other difficult tracking conditions.

Initialization: When an object is detected in one camera, we try to initialize it in the other cameras. For this purpose, the epipolar lines are calculated that correspond to the estimated target location. Samples are then placed stochastically around these lines and the velocity components are chosen from Gaussian distributions. If the object is already visible in more than one camera, the intersection of the corresponding epipolar lines is calculated and the samples are distributed around this point.

Reinitialization: If less than $b \cdot N$ of the samples fulfill the appearance condition which is explained in Eq. 3, we consider the object to be lost. In this case, we use the epipolar lines and their intersections to reinitialize an object during tracking. A fraction of the samples are then spread around the epipolar lines of the other cameras while the remaining samples are propagated normally.

5 Best View Selection

Transmission of presentations in intelligent environments are attractive due to wide application areas such as surveillance or training tasks like a virtual classroom. In the context of the ViRoom, we are interested in a front view of the face whereby the person is moving freely in our smart room. Accordingly, we have developed a automated virtual editor which creates a video stream by switching to the best view in the room. Given several camera inputs, it automatically chooses the one that gives the best front view of the target. The camera hand-over is controlled on the basis of the tracking results by evaluating the Bhattacharyya coefficient (see Eq. 2) of the mean state of each tracker

$$\rho[p_{E[S]}, q_f] = \sum_{u=1}^m \sqrt{p_{E[S]}^{(u)} q_f^{(u)}} \quad (4)$$

$$E[S] = \sum_{n=1}^N \pi^{(n)} \mathbf{s}^{(n)}. \quad (5)$$



Fig. 3. The best view selection according to the Bhattacharyya coefficients of the individual trackers is shown. The small images on the bottom show the current tracking results of the individual camera views as white ellipses and the corresponding Bhattacharyya coefficients. In the top row the best target model and the output of the virtual editor are displayed.

As the Bhattacharyya coefficient represents a similarity measure in respect to the target histogram, the virtual editor chooses always the camera view which provides the highest Bhattacharyya coefficient in means of the facial view. Figure 3 illustrates the best view selection on the basis of the individual Bhattacharyya coefficients.

6 Results

In this Section, experimental results demonstrate the capabilities and limitations of our distributed tracking system. All images are captured from live video streams and have a size of 160×120 pixels. The application runs at 4-5 frames per seconds — without any special optimization — on Pentium II PCs at 1GHz under Linux where each of the three cameras is attached to its own computer.

The capability of the virtual editor is illustrated in Figure 4. It can be seen that the camera hand-over automatically chooses the best front view of the tracked face even if it is partly occluded.

The initialization of the tracker plays an important role in the multi-camera tracking. However, when there are several possible objects in the neighborhood of the epipolar lines, a wrong object can be selected. Such a scene is shown in Figure 5. In the top row, the situation is handled correctly as the target is occluded in the middle camera and not initialized. In the second row, the target is not localized correctly whereas in the third row a wrong target is selected. In both cases, the target is occluded in two cameras, so that the samples are spread along an epipolar line and not around a intersection point.



Fig. 4. The virtual editor automatically chooses the best front view of the tracked face.

7 Conclusion

The proposed system describes a step towards intelligent processing of visual data. An example setup is presented with multiple cameras and a tracker for each of them. To track a target robustly in all camera views, multiple models are recorded and the trackers select automatically the model that matches best. By implementing a virtual editor, an on-line documentation is produced on the basis of tracking results and camera control.

Information exchange between the individual trackers is currently only used for the (re)initialization process by applying epipolar geometry. We will integrate the use of a common state for all trackers based on 3D information. Furthermore,



Fig. 5. The initialization step can cause problems in tracking, if there are several equally good candidates in the vicinity of the epipolar lines.

we will enhance the virtual editor. The camera selection should be made in such a way that the resulting stream is pleasant to watch and accordingly, short movie cuts should be avoided. In addition, the Bhattacharyya coefficient which is the basis for the best view selection will be combined with alternative decision rules. We have also planned to compute virtual views from the available images to increase the information content of the video stream. A teleconferencing setup with multiple people is another interesting extension. The virtual editor should be able to locate the person which is the main actor within the observed scene.

Acknowledgment

The authors acknowledge the support by the European Commission project STAR (IST-2000-28764). We thank Petr Doubek and Stefaan De Roeck for the multi-camera set-up and calibration, and Bart Vanluyten and Stijn Wuyts for including the active cameras.

References

1. F. Aherne, N. Thacker and P. Rockett, *The Bhattacharyya Metric as an Absolute Similarity Measure for Frequency Coded Data*, Kybernetika, pp. 1-7, Vol. 32(4), 1997.
2. C.J.C. Burges, *A tutorial on support vector machines for pattern recognition*, Data Mining and Knowledge Discovery, Vol.2(2): pp. 955-974, 1998.
3. D. Comaniciu, F. Berton and V. Ramesh, *Adaptive Resolution System for Distributed Surveillance*, Real-Time Imaging, pp. 427-437, Vol. 8, 2002.
4. D. Comaniciu, V. Ramesh and P. Meer, *Real-Time Tracking of Non-Rigid Objects using Mean Shift*, CVPR, pp. 142-149, Vol. 2, 2000.
5. P. Doubek, T. Svoboda and L. Van Gool, *Monkeys - a Software Architecture for ViRoom - Low-Cost Multicamera System*, ICVS, pp. 386-395, 2003.
6. M. Isard and A. Blake, *CONDENSATION - Conditional Density Propagation for Visual Tracking*, International Journal on Computer Vision, pp. 5-28, Vol. 1(29), 1998.
7. T. Kailath, *The Divergence and Bhattacharyya Distance Measures in Signal Selection*, IEEE Transactions on Communication Technology, COM-15(1) pp. 52-60, 1967.
8. S. Kahn, O. Javed and M. Shah, *Tracking in Uncalibrated Cameras with Overlapping Field of View*, PETS, 2001.
9. J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale and S. Shafer, *Multi-Camera Multi-Person Tracking for EasyLiving*, International Workshop on Visual Surveillance, pp. 3-10, 2000.
10. K. Nummiaro, E. Koller-Meier and L. Van Gool, *An Adaptive Color-Based Particle Filter*, Journal of Image and Vision Computing, pp. 99-110, Vol 21(1), 2003.
11. P. Pérez, C. Hue, J. Vermaak and M. Gangnet, *Color-Based Probabilistic Tracking*, ECCV, pp. 661-675, 2002.
12. T. Svoboda, H. Hug and L. Van Gool, *ViRoom - Low Cost Synchronised Multicamera System and its Self-Calibration*, DAGM, pp. 515-522, 2002.
13. M.M. Trivedi, I. Mikic and S.K. Bhonsle, *Active Camera Networks and Semantic Event Databases for Intelligent Environments* Proceedings of the IEEE Workshop on Human Modelling, Analysis and Synthesis, 2000.